A Computational Theory of Adaptive Behavior Based on an Evolutionary Reinforcement Mechanism

J. J McDowell
Department of Psychology
Emory University
Atlanta, GA 30322
jack.mcdowell@emory.edu

Paul L. Soto
Department of Psychology
University of Florida
Gainesville, FL 32611
psoto@ufl.edu

Jesse Dallery
Department of Psychology
University of Florida
Gainesville, FL 32611
jdallery@ufl.edu

Saule Kulubekova
Department of Psychology
Emory University
Atlanta, GA 30322
skulube@emory.edu

ABSTRACT

Two mathematical and two computational theories from the field of human and animal learning are combined to produce a more general theory of adaptive behavior. The cornerstone of this theory is an evolutionary algorithm for reinforcement learning that instantiates the idea that behavior evolves in response to selection pressure from the environment in the form of reinforcement. The evolutionary reinforcement algorithm, along with its associated equilibrium theory, are combined with a mathematical theory of conditioned reinforcement and a computational theory of associative learning that together solve the problem of credit assignment in a biologically plausible way. The result is a biologically-inspired computational theory that enables an artificial organism to adapt continuously to changing environmental conditions and to generate adaptive state-action sequences.

Track: Artificial Life, Evolutionary Robotics, Adaptive Behavior

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.6 [**Simulation and Modeling**]; J.4 [**Social and Behavioral Sciences**] – *Psychology*.

General Terms: Algorithms, Theory.

Keywords: Evolutionary algorithms, reinforcement learning, adaptive behavior, adaptive agents, conditioned reinforcement, credit assignment, stimulus control, matching theory, delay-reduction theory, Rescorla-Wagner rule.

1. INTRODUCTION

Learning may be viewed alternatively as acquiring knowledge or as behaving adaptively. The former view, which is common in artificial life and machine learning [12], leads to a focus on how and what rational agents learn about their environments. The latter view leads to a focus on how agents adjust their behavior to prevailing environmental conditions.

The view of learning as behaving adaptively characterizes an approach to human and animal learning known as the experimental analysis of behavior [11]. Specific mathematical and computational theories in this field have advanced to an extent that they now can inform research on artificial life in useful ways. The purpose of this article is to combine four specific behavior theories into a more comprehensive computational theory of adaptive behavior. The centerpiece of this application of behavior theory is a computational reinforcement mechanism that is based on evolutionary principles [8]. We combine this mechanism, and its associated equilibrium theory, with two additional theories that solve the problem of credit assignment in a biologically plausible way. The additional theories are Mazur's [6] hyperbolic delay theory of conditioned reinforcement and the Rescorla-Wagner theory of associative learning [1, 14]. The result is a biologically-inspired computational theory that allows an artificial organism to adapt continuously to changing environmental conditions, and to produce adaptive state-action sequences.

In the next section the reinforcement mechanism will be described, computational experiments testing the ability of the mechanism to generate behavior typical of live organisms will be summarized, and the mechanism's ability to produce continuously adaptive behavior in an artificial organism will be illustrated. Following this, the reinforcement mechanism will be combined with the Mazur and the Rescorla-Wagner theories to produce a more comprehensive computational theory of adaptive behavior.

2. THE REINFORCEMENT MECHANISM

The reinforcement mechanism is based on the idea that behavior evolves in response to selection pressure exerted by the environment in the form of reinforcement. The algorithm consists of a fitness rule, and rules of selection, reproduction, and mutation that govern the evolution of a population of potential behaviors in an agent, or artificial organism. At each moment, or time tick, a new generation of potential behaviors is produced and one behavior from the population is emitted. This generates a constant stream of behavior over time. As will be explained in more detail in the next section, selection pressure exerted by the environment operates in such a way that relatively successful behaviors, which are behaviors that either produce reinforcement or are similar to those that produce reinforcement, tend to become more frequent in the population of potential behaviors, whereas relatively unsuccessful behaviors tend to become less frequent. It is important to keep in mind that in this theory, evolution operates on a population of potential behaviors in a single agent, not on a population of agents. The specific details of the behavioral repertoire and of the evolutionary rules are described next.

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-59593-186-4/06/0007…$5.00.

2.1 Organism and Algorithm

Let an artificial organism consist of a repertoire, or population, of 100 potential behaviors, each of which is defined by an integer ranging from 0 through 1,023. The top panel of Figure 1 shows a repertoire of 100 behaviors selected at random from the permissible range, and arranged in arbitrary order along the *x*-axis. The integers that represent the behaviors may be thought of as the behaviors' phenotypes, and the behaviors may be sorted into classes based on these phenotypes. For the present discussion we will consider just two classes, a target class that consists of phenotypes 0 through 40, and an extraneous class that consists of all other phenotypes. At each moment, or tick, of time the artificial organism emits a behavior from one of these classes, with probabilities equal to the relative frequencies of the phenotypes in the classes. For example, if the repertoire at a particular time tick consists of 20 behaviors that fall in the target class and 80 behaviors that fall in the extraneous class, then the probability that a behavior will be emitted from the target class at that moment is 20/100 or 0.20.

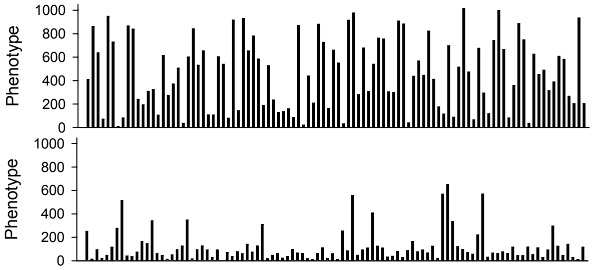

Figure 1. Top: Representation of a random initial repertoire consisting of 100 behaviors with integer phenotypes ranging from 0 through 1,023. Bottom: Repertoire after 433 generations during which phenotypes 0 through 40 were occasionally reinforced. In both panels the behaviors are arranged in arbitrary order along the *x*-axis.

Suppose the organism is placed in an environment in which the emission of a behavior from the target class occasionally produces reinforcement, unpredictably, but with a known average rate, *r*. Emissions of behavior from the extraneous class are never reinforced. In behavior analysis this environment is said to arrange a random interval (RI) schedule of reinforcement. Reinforcing a target behavior identifies it as successful with respect to other behaviors in the repertoire. The computational theory uses this information to produce the next generation of potential behaviors by choosing "parent" behaviors from the population on the basis of their similarity to the reinforced behavior. Behaviors that are more similar to the reinforced behavior are more likely to be chosen as parents. These parents then produce the next generation of "child" behaviors. Selecting specific parent behaviors entails first calculating a fitness value for every behavior in the repertoire. This value represents the similarity of a specific behavior to the reinforced behavior. One way to define fitness is as the absolute value of the difference between a behavior's phenotype and the phenotype at the midpoint of the target class. Defined in this way, smaller fitness values are associated with fitter behaviors, which are behaviors that are more like the reinforced behavior. Given fitness values for every behavior in the population, individual parent behaviors are selected using a probability density function that assigns higher probabilities of being selected to fitter behaviors, and lower probabilities of being selected to less fit behaviors.

Once parent behaviors are chosen, they combine to produce child behaviors. If the integer value that defines a behavior is its phenotype, then the binary representation of this integer can be considered its genotype. Bits from binary representations of two parent behaviors can be combined in various ways to produce child behaviors. The child behaviors produced in this way then replace some percentage of the initial population of behaviors.

After the new generation of behaviors is produced, a small amount of random mutation is added to the population by, for example, flipping a bit in the binary representation of a small number of randomly selected behaviors. The 100 potential behaviors in the new repertoire are then sorted into target and extraneous classes and, based on their relative frequencies as described earlier, a specific behavior is emitted from one of the classes. If the behavior is from the target class and it is reinforced, then the process of calculating fitness, selecting parents, producing a new population of behaviors, and adding random mutation is repeated. If the behavior is from the target class but is not reinforced, or if it is from the extraneous class, then parents selected at random produce the next generation, and a small amount of random mutation is added to the population of behaviors.

This algorithm consists of a fitness rule, and rules of selection, reproduction, and mutation. The fitness rule specifies how the similarity between a specific behavior and the reinforced behavior is represented. The selection rule is constituted by the form and parameter values of the parental-selection density function. The reproduction rule specifies how parent behaviors combine to produce child behaviors. And the mutation rule specifies how and how many mutants are produced. Each of these rules can be implemented in various ways. For example, fitness can be calculated with respect to various properties of the target class. The parental selection function must assign higher probabilities of being selected to fitter behaviors, but its form could be linear, uniform, or exponential, among other possible forms. There is a similar variety of ways to implement the reproduction and mutation rules. To anticipate the results of computational experiments that will be summarized in the next section, it evidently does not matter how these rules are implemented, only *that* they are implemented. Within wide ranges, any set of fitness, selection, reproduction, and mutation rules generates the same pattern of equilibrium states.

2.2 Computational Studies of the Algorithm

The general effect of the evolutionary algorithm is illustrated in Figure 1. Beginning with a random initial repertoire (top panel), a target behavior (phenotypes 0 through 40) was occasionally reinforced. The bottom panel shows the repertoire after 433 generations. Overall, the population of behaviors is now much fitter, that is, the phenotypes of the individual behaviors are closer to the midpoint of the target class, which is 20.

When the evolutionary algorithm is run for many generations, the behavior of the artificial organism eventually reaches an equilibrium such that the momentary rate of behaviors from the target class varies around some mean value, *R*. Reinforcement tends to pull the population of behaviors into the target class while non-reinforcement and mutation tend to return the population of behaviors to the baseline distribution of target and extraneous behaviors.

The behavior of many species of live organism also reaches an equilibrium state that varies around some mean rate, R, in environments that arrange occasional reinforcement for a target behavior. To evaluate the evolutionary algorithm as a theory of live behavior dynamics, it is necessary to summarize briefly what is known about the equilibrium states of behavior produced by live organisms.

The relationship between R and r for live organisms has been studied extensively, and is known to be described by the hyperbolic equation,

$$R = \frac{kr}{r + r_e}, \qquad (1)$$

which expresses the steady-state rate of a specific target behavior, R, as a function of the steady-state rate of reinforcement obtained for the target behavior, r [3]. The quantities, k and r_e, are parameters of the hyperbola. This equation is part of a very successful family of equations known as matching theory [2, 9], which describes how steady-state behavior is related to environmental conditions such as the rate of reinforcement obtained for the behavior. Equation 1 is the fundamental equation of matching theory. According to the theory, the parameters, k and r_e, in Equation 1 are interpreted respectively as the maximum rate of the target behavior, and the aggregate rate of background reinforcement, which is reinforcement delivered for behaviors other than the target behavior, or delivered for free.

Equation 1 has been extensively verified in dozens of experiments with many species, including humans, and many types of behavior and reinforcement [2]. It is now generally accepted as a fundamental, quantitatively accurate, statement of how reinforcement governs the behavior of biological organisms in the steady state. In an environment that arranges an average reinforcement rate, r, for a particular behavior, a human or animal will eventually come to emit that behavior at the average rate, R, specified by Equation 1, and will continue to do so until r changes. It is important to recognize that in most experimental environments, many instances of the target behavior are emitted but only a few actually produce reinforcement. Hence, Equation 1 is readily applicable to many naturalistic human and animal environments [7].

The dynamics that specify how the steady state described by Equation 1 is achieved are not well understood. Much research has been devoted to this topic [e.g., 12], but none of the proposed solutions has been widely accepted. The evolutionary theory described here is a recent attempt to solve this problem. To be considered a reasonable candidate for a dynamic theory of behavior, its equilibrium states must be described by Equation 1.

Computational studies have been conducted to investigate this requirement, and to determine whether the theory's pattern of equilibrium states depends on the details of the algorithm's implementation. To investigate the first issue, equilibria generated by the evolutionary algorithm over a wide range of average reinforcement rates, r, were analyzed [8]. Values of R in these environments were averages over 5,000 to 45,000 generations, which produced very small standard errors. Equation 1 was then fitted to the r-R data pairs by the method of least squares, and the residuals were examined for randomness. For 57 sets of equilibria, each of which consisted of 9 to 11 r-R data pairs, Equation 1 accounted for, on average, 99% of the variance of the Rs (range: 68% to 100%). Statistical tests of the residuals detected deviations from randomness in 12 (21%) of the 57 fits. These results indicated that Equation 1 provided a good description of the equilibrium states produced by the evolutionary theory.

It is possible that function forms different from Equation 1, but with similar differential properties, might describe these equilibria just as well. To test this possibility, an asymptotic exponential, an asymptotic power function, and a ramp function, each having two parameters, were fitted to the 57 sets of equilibria generated by the evolutionary algorithm [8]. On average, the comparison forms accounted for, respectively, 96%, 91%, and 85% of the variance of the Rs. The differences between each of these percentages and the percentage of variance accounted for by Equation 1 (99%) were statistically significant. In addition, statistical tests of randomness applied to the residuals detected deviations from randomness in 44 (77%), 42 (74%), and 51 (89%) of the 57 fits for the three comparison forms respectively. These results indicated that the equilibrium states produced by the evolutionary mechanism were *specifically* described by a hyperbola (Equation 1), as required by matching theory.

It is possible that the hyperbolic form of the equilibrium states generated by the evolutionary algorithm depended critically on a specific feature of the algorithm. Of particular concern is the form of the parental selection function. The possibility that this form somehow determines the form of the equilibrium states was tested in experiments using linear, uniform, and exponential probability density functions to select parents for mating. All three forms were found to produce equilibrium states that were accurately and specifically described by Equation 1. In addition to different forms of the parental selection functions, two distinct methods of defining fitness, two methods of combining bits to produce child behaviors, two methods of mutation, and various combinations of these methods were also studied. In every case, the resulting equilibrium states were accurately and specifically described by Equation 1 [8].

Taken together, these computational experiments showed that Equation 1, which is known to describe the behavior of live organisms, is a unique and robust emergent property of an evolutionary behavior dynamics. The evolutionary algorithm, then, appears to be a viable candidate for a theory of behavior dynamics in live organisms.

2.3 Adaptive Behavioral Dynamics

Live organisms are able to continuously adapt to changing environmental conditions. The evolutionary algorithm equips artificial organisms with a similar ability. One way to illustrate this is to change the selection pressure during an experimental run and observe how the artificial organism responds. The top panel of Figure 2 shows, for two experimental runs, the momentary probability of emission of a target behavior that had a baseline probability of 0.04. Recall that this momentary probability is just the relative frequency of potential behaviors that fall in the target class. The target behavior was occasionally reinforced on an RI schedule during 6000 generations, or time ticks. As explained earlier, an RI schedule arranges reinforcement unpredictably, but with a known mean rate.

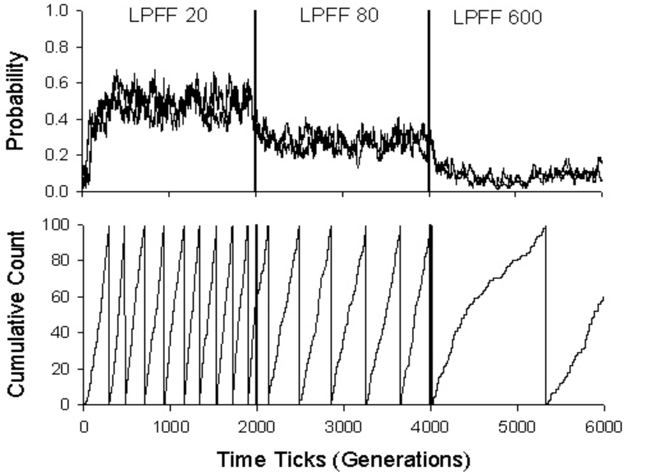

Figure 2. Momentary probabilities (top panel) and cumulative counts (bottom panel) of a target behavior that was occasionally reinforced on an RI schedule. Two experimental runs are plotted in the top panel; the cumulative counts from one of the runs are plotted in the bottom panel. The mean of the linear parental selection function (LPFF) was changed from an initial value of 20 to a value of 80 at generation 2001, and then to a value of 600 at generation 4001.

During the experimental runs, a one-parameter, linear parental-selection density function was used, namely,

$$p(x) = -\frac{2}{9\mu^2}x + \frac{2}{3\mu}, \tag{2}$$

for $0 \le x \le 3\mu$, where the single parameter, μ, represented the mean of the density function. In this equation x represents fitness and *p(x)* is the probability density associated with a behavior of fitness, x, being chosen as a parent. This is the simplest linear probability density function that depends only on its mean. When the mean is small, only highly fit behaviors can be selected as parents, and hence the selection pressure is strong; when the mean is large, less fit behaviors can be selected as parents and hence the selection pressure is relatively weak.

During the first 2000 generations of the experimental runs shown in the top panel of Figure 2, the mean of the linear parental selection function (LPFF) was 20. The selection pressure exerted by this density function was reduced at generation 2001 by increasing the mean of the function to 80, and it was reduced further at generation 4001 by increasing the mean to 600. As shown in the figure, the momentary probability of the target behavior increased over the first 300 or 400 generations from its baseline level to a roughly stable probability of about 0.5. When the selection pressure was reduced at generation 2001, the momentary probability of the target behavior fell over the course of about 300 generations and reached a new equilibrium of approximately 0.25. A similar decline and then stabilization of the momentary probability of emission of the target behavior occurred when the selection pressure was reduced further at generation 4001.

The translation of the momentary probabilities into behaviors is shown in the bottom panel of Figure 2 for one of the experimental runs. The cumulative number of target behaviors is plotted against time. Whenever the cumulative number reached 100 it was reset to zero. One advantage of a cumulative plot of behavior is that its slope gives the behavior's time rate of occurrence. The cumulative plot in Figure 2 shows that the target behavior was emitted at a roughly constant momentary rate, which decreased with decreasing selection pressure. Roughly constant momentary rates of behavior are widely observed in experiments with live organisms on RI schedules [11]. The artificial organism's adaptive response to changes in selection pressure is not limited to the linear parental selection function used in this example, or to any other detail of the evolutionary algorithm. Every instantiation of the algorithm evidently exhibits this property.

In addition to enabling an artificial organism to change the rate of a targeted behavior in response to changes in selection pressure, the evolutionary mechanism also enables an artificial organism to respond adaptively when the environmental requirements change so as to favor a different behavior. Figure 3 shows one experimental run of an artificial organism with two classes of behavior having baseline probabilities of 0.04 (Behaviors 1 and 2), and two classes of behavior having baseline probabilities of 0.46. During the first 2000 generations of the experimental run, Behavior 1 (solid line in the figure) was occasionally reinforced on an RI schedule. At generation 2001, Behavior 1 was put on extinction, that is, reinforcement was withdrawn, and Behavior 2 was occasionally reinforced using the same RI schedule. At generation 4001 the original conditions were reinstated.

The momentary probabilities in Figure 3 show that the probability of emission of Behavior 1 increased over the course of about 400 generations from its initial low value to a roughly stable probability of about 0.4. When reinforcement was withdrawn at generation 2001, the probability of Behavior 1 decreased over about 300 generations to its initial low value, and when reinforcement was resumed at generation 4001 the probability of Behavior 1 again increased to a steady-state value of about 0.4. Similar changes were evident for Behavior 2, beginning with a low initial probability, followed by a transition to a higher equilibrium probability when the behavior was reinforced, followed by a return to the lower probability when reinforcement was withdrawn.

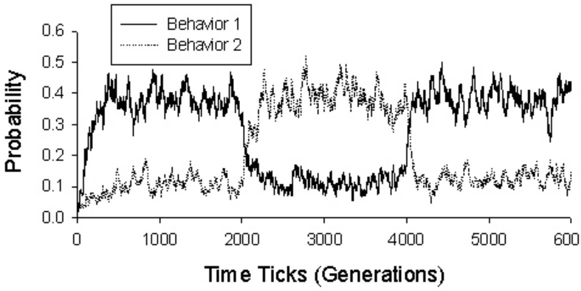

Figure 3. Momentary probabilities of two behaviors for one experimental run where Behavior 1 (solid line) was occasionally reinforced on an RI schedule during generations 0 through 2000 and generations 4001 through 6000, and Behavior 2 (dotted line) was occasionally reinforced on an RI schedule during generations 2001 through 4000.

Other examples of the artificial organism's adaptive behavior could be adduced. For example, smooth transitions from one equilibrium condition to another occur when the mean rate of reinforcement, *r*, is changed. In fact, the organism responds with more or less smooth transitions between equilibria when any detectable feature of the environment changes, including qualitative changes such as the form of the parental selection function.

3. THE GENERAL THEORY

The approach to reinforcement learning presented here is different from approaches that are commonly found in the machine learning and artificial intelligence literature. More familiar approaches include those based on the expected utility or value of different courses of action [4, 14], and those focused on finding, often by using evolutionary algorithms, the best action or policy in particular sets of circumstances [10].

One of the distinguishing features of the evolutionary mechanism described in this article is that it does not entail a performance criterion, and in that sense the organism it animates is not a rational agent. In utility-based reinforcement learning, maximization of expected utility is the performance criterion. In action-based reinforcement learning, policies that optimize some benefit to the organism are typically sought. In contrast, our evolutionary mechanism operates without information about the longer term or overall benefits associated with different courses of action. Like organic evolution, which is not an optimizing process, behavioral evolution simply acts from moment to moment in a way that causes behavior to change in response to changes in environmental conditions.

A second distinguishing feature of our approach is that learning occurs continuously in time rather than on a trial-by-trial basis. Consider, for example, a rat in a T-maze that is learning to turn left or right at the end of a runway. After making a turn, the rat is put back in the start box, has another go at the maze, and so on. Many reinforcement learning algorithms follow this paradigm. In contrast, consider a rat in an experimental chamber with a lever protruding from one wall. The rat may press the lever at any time, or explore the chamber, or rear, or preen, and so on. Consequences that are arranged for lever pressing must control that behavior in real time and in the context of other behaviors. Our approach models this more naturalistic, “free-behavior”, situation.

3.1 Mechanics of Conditioned Reinforcement

Combining the evolutionary reinforcement mechanism and its associated equilibrium theory (Equation 1) with the Mazur and the Rescorla-Wagner theories produces a computational account that is applicable to more complicated situations. For example, learning algorithms in artificial life research often deal with state-action sequences [13]. An agent in a given state has a set of possible actions, each of which puts it in a new state with a new set of possible actions, and so on. Consider, for example, a grid world with an agent in one state (i.e., a box in the grid) and a reinforcer available in another state. The task might be to move through the grid world to get to the reinforcer, perhaps in the fewest number of steps. Each trip through the grid world constitutes a series of state-action sequences, or a policy. Once the reinforcer is encountered, credit must be assigned to each step in the trip, or to the policy as a whole, to make that specific trip more or less likely to occur in the future. The agent is then put back into the grid world, takes another trip, and so on.

A state in artificial learning is comparable to a discriminative stimulus (usually abbreviated S^D) in the experimental analysis of behavior. A discriminative stimulus may signal the opportunity to obtain reinforcement, or it may signal the opportunity to obtain a new discriminative stimulus, that is, to change state. For example, in the presence of a 1000-Hz tone, a particular behavior may occasionally turn on a yellow light, and then in the presence of the yellow light, a particular behavior may occasionally produce reinforcement. In this example, the tone signals the opportunity to change state, and the yellow light signals the opportunity to obtain reinforcement. In the experimental analysis of behavior, this sequence of discriminative stimuli and their signaled behaviors is referred to as a chained schedule of reinforcement. Each discriminative stimulus identifies one link in the chain. A chained schedule with three links is illustrated in Figure 4.

A trip through a grid world is comparable to an organism behaving on a chained schedule of reinforcement. In the presence of an initial discriminative stimulus or state, one class of behavior turns on a new discriminative stimulus, that is, causes a change of state. A class of behavior in the presence of this new discriminative stimulus turns on yet another discriminative stimulus, and so on, until one class of behavior in the presence of a final discriminative stimulus produces reinforcement. This is a sequence of state-action mappings that ends with reinforcement.

The evolutionary reinforcement mechanism described in this article can be applied to a chained schedule by associating a distinct repertoire of behaviors with each discriminative stimulus. Each repertoire evolves only in the presence of its discriminative stimulus. This association of distinct repertoires with different discriminative stimuli is consistent with the known effects of discriminative stimuli on the behavior of live organisms [11].

Consider now the chained schedule illustrated in Figure 4, beginning with the final link (Link 1) in the chain. In the presence of S^D_1, the organism can emit behavior from, say, three classes, which are represented by arrows in the figure. Suppose behavior from one of the classes, the target class (thick arrow), produces reinforcement on a random interval schedule. As long as S^D_1 is present, the repertoire of behavior associated with it evolves according to the evolutionary algorithm. Now suppose we turn on S^D_2. In the presence of this discriminative stimulus, the organism again can emit behavior from three classes, represented by arrows in the figure, and behavior from one of the classes turns on S^D_1 according to a random interval schedule. As long as S^D_2 is present, its associated repertoire of behavior evolves according to the evolutionary algorithm. But what provides the selection pressure in this link of the chain? In the case of live organisms, the onset of S^D_1, the discriminative stimulus that signals reinforcement, provides the selection pressure.

We know from research with live organisms that discriminative stimuli that are associated with reinforcement themselves acquire reinforcing properties. In other words, they become conditioned reinforcers, which organisms will work to produce. Specifically, Mazur [6] showed that the steady-state reinforcing value, *V*, of a discriminative stimulus is a hyperbolic decay function of the time between its onset and the moment of reinforcer delivery,

$$V = \frac{a}{1 + bx}, \tag{3}$$

where *x* is the latency of reinforcement following the onset of the discriminative stimulus, and *a* and *b* are parameters of the equation. The longer the delay to reinforcement after the onset of the discriminative stimulus, the lower is its reinforcing value. Equation 3 applies to simple situations where one reinforcement latency is repeatedly associated with a discriminative stimulus. On a random interval schedule, however, many latencies are associated with the discriminative stimulus because reinforcement occurs at unpredictable times. The onset of each latency is the moment that responding resumes after reinforcement. Mazur [5] showed that in the case of many latencies, the overall or total reinforcing value of a discriminative stimulus, V_T, is the average of the reinforcing values of the individual latencies, V_i,

$$V_T = \sum_{i=1}^{n} p_i V_i , \tag{4}$$

where p_i is the probability of the i^{th} latency. But on a random interval schedule, latencies are exponentially distributed such that that the probability of a latency of magnitude, x, is

$$p(x) = re^{-rx} , \tag{5}$$

where r is the mean rate of reinforcement arranged by the schedule. It follows from Equation 4 that the overall or total reinforcing value of a discriminative stimulus associated with a random interval schedule is

$$V_T = \int_0^\infty re^{-rx} V dx , \tag{6}$$

which evaluates to

$$V_T = a(r/b)e^{r/b}\Gamma(0, r/b). \tag{7}$$

Equation 7 expresses the steady-state or equilibrium reinforcing value of a discriminative stimulus associated with a random interval schedule, as a function of the average rate of reinforcement, r, delivered by the schedule, and the fixed Mazur parameters, a and b. This equation asserts that the reinforcing value of the discriminative stimulus is a monotonically increasing function of r that approaches an asymptote, a, with rapidity governed by b.

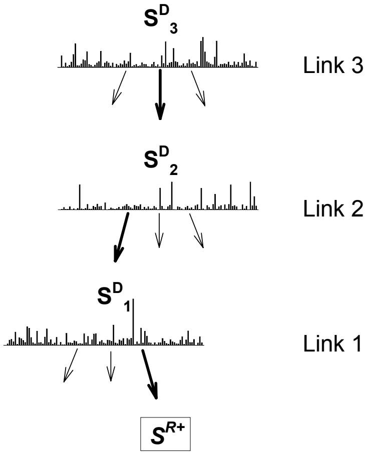

Figure 4. A chained schedule of reinforcement with three links. Each link is identified by a different discriminative stimulus, S^D. In the presence of an S^D, one class of behavior (thick arrows) turns on the next S^D, or produces reinforcement (S^{R+}). This is a sequence of state-action mappings that ends with reinforcement. A separate repertoire of behaviors is associated with each discriminative stimulus, and evolves only in the presence of that S^D.

According to our extension of Mazur's theory, Equation 7 must give the reinforcing value of S^D_1 in the chained schedule illustrated in Figure 4 *at equilibrium*, that is, after sufficient exposure to the schedule. But how does this equilibrium develop? We propose that discriminative stimuli acquire reinforcing value through a process of associative learning that can be described by the well-known Rescorla-Wagner theory. This theory has been applied successfully to associative learning in both live and artificial organisms [1, 14]. To apply the theory to discriminative stimuli, the associative strength or reinforcing value, V, of a discriminative stimulus at time step, t, is written as a function of its associative strength on the previous time step, t-1, its salience, α, the salience of the reinforcing stimulus, β_λ, and the ultimate level of associative learning that the reinforcing stimulus will support, λ,:

$$V_t = V_{t-1} + \alpha\beta_\lambda\left(\lambda - V_{t-1}\right). \tag{8}$$

This equation describes the momentary reinforcing value of a discriminative stimulus which, according to the equation, is principally a function of its reinforcing value on the previous step and the salience of the stimulus and of the reinforcer. As examples of salience, a dim light is probably a less salient discriminative stimulus than a bright light; food reinforcement is likely to be more salient when an organism is food deprived than when it is not. Lambda, the maximum reinforcing value, is conventionally assigned a numeric value of 1 when reinforcement occurs on a time step and 0 when it does not. This is because no reinforcing value can accrue to the discriminative stimulus when reinforcement does not occur. The salience parameters, α and β, are typically set to values between 0 and 1, and the value of β may differ when reinforcement occurs than when it does not because the former condition is likely to be more salient than the latter. Using these conventions, the reinforcing value, V, of the discriminative stimulus varies between 0 and 1. At each time step V is incremented if reinforcement occurs and decremented (because $\lambda = 0$) if it does not.

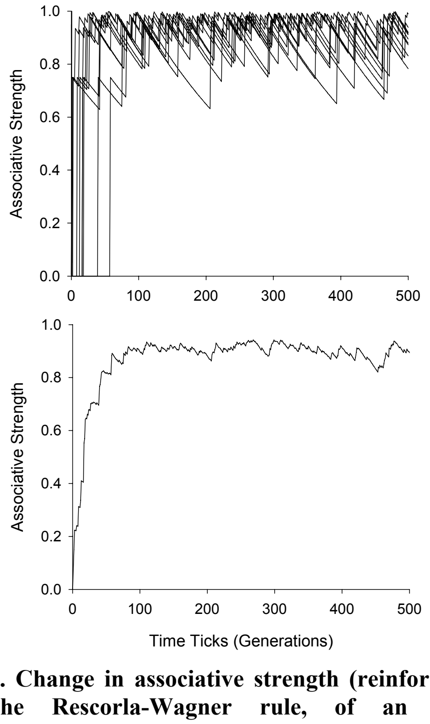

Figure 5. Change in associative strength (reinforcing value), using the Rescorla-Wagner rule, of an S^D where reinforcement occurs on a random 5% of the time steps. Top: Ten experimental runs with α = 0.75, β_1 = 1 and β_0 = 0.006. Bottom: Average of the 10 runs.

The acquisition of reinforcing value by a discriminative stimulus on an RI schedule can be illustrated by applying Equation 8 to experimental runs where reinforcement occurs on, say, 5% of the time steps. Ten such experimental runs are shown in the top panel of Figure 5, for which $\alpha = 0.75$, $\beta_1 = 1$ and $\beta_0 = 0.006$. The salience of reinforcer absence (β_0) is much smaller than the salience of its presence (β_1) because the reinforcer is absent 95% of the time. The average associative strength for the ten runs is plotted in the bottom panel of the figure. These plots show that the associative strength of a discriminative stimulus under these conditions increases to a final value over about 100 generations, and then varies around it. This is an example from ongoing research in our laboratories, which indicates that on RI schedules, Equation 8 produces stable reinforcing-value equilibria that are directly proportional to the average reinforcement rates arranged by the schedules. This outcome is qualitatively consistent with our extension of the Mazur theory, Equation 7. Studies of the quantitative agreement between Equation 7 and the reinforcing-value equilibria generated by Equation 8 are currently underway.

Notice that Equations 7 and 8 are theories of the statics and dynamics of conditioned reinforcement, just as Equation 1 and the evolutionary reinforcement algorithm are theories of the statics and dynamics of reinforcement learning. In our more general theory, the mechanics (that is, the statics and dynamics) of reinforcement learning governs behavior in each link of the chained schedule illustrated in Figure 4. The mechanics of conditioned reinforcement governs how the changes of state in links two and three acquire reinforcing value. According to this mechanics, the reinforcing value of $S^D{}_1$ as a consequence for behavior in the second link of the chain depends on its association with the reinforcer in the first link, and the reinforcing value of $S^D{}_2$ as a consequence for behavior in the third link of the chain depends on its association with $S^D{}_1$ in the second link. This theory of conditioned reinforcement is a solution to the problem of credit assignment that is consistent with the known effects of conditioned reinforcement in live organisms. There are other ways to model chains of behavior. For example, Touretzky and Saksida [15] have discussed a knowledge-based method and have applied it to an experimental task known as delayed matching to sample.

3.2 The Synthesis

The last piece of our theory is a connection between the mechanics of conditioned reinforcement and the mechanics of reinforcement learning. We must specify how the conditioned reinforcing values of the discriminative stimuli that are produced by behavior in the second and third links of the chain affect reinforcement learning in those links. To accomplish this, we need to know how reinforcing value is represented in the evolutionary theory in the first place. McDowell [8] proposed that the value, or magnitude, of a reinforcer is represented in the theory by the mean of the parental selection function. Recall that a parental selection function is used to select parent behaviors for mating based on their similarity to the reinforced behavior. As one example, Equation 2 is a linear parental selection function that depends only on its mean. It is possible to construct single-parameter density functions of any form that depend only on their means [8]. As illustrated in Figure 2, the mean of the function determines the selection pressure exerted by the reinforcer. When the mean is small, very fit behaviors are likely to be chosen as parents and hence the selection pressure is relatively strong; when the mean is large, less fit behaviors can be chosen as parents and hence the selection pressure is relatively weak. Notice that changes in the momentary probability of the target behavior illustrated in Figure 2 occur even though the rate of reinforcement is constant. The rate of reinforcement determines how often a selection event occurs; the mean of the parental selection function determines the impact of each selection event. Both variables affect the overall fitness of the population of potential behaviors, and hence the rate of the target behavior.

If the mean of the parental selection function represents reinforcing value, then as reinforcing value accrues to a discriminative stimulus in a link of a chained schedule, the mean of the parental selection function that entails that stimulus as a consequence must decrease. The simplest way to represent this is with the reciprocal function,

$$\mu_i = \frac{\mu_{i-1}}{V_{i-1}}, \tag{9}$$

where μ represents the mean of the parental selection function operating in a link, and V represents the reinforcing value of the discriminative stimulus associated with a link. The subscripts identify the link to which the quantities refer. Recall that the numerical value of V is conventionally taken to vary from 0 to 1. Using this convention, and the notation of Equation 9, μ_0 represents the mean of the parental selection function for the terminal reinforcer, the value (V_0) of which is 1. To understand Equation 9, consider as an example, μ_2, the mean of the parental selection function used in the second link of the chained schedule illustrated in Figure 4. This mean is updated *during the first link of the schedule* according to Equation 9. This is because during the first link, μ_1 is constant, while V_1 is updated according to the Rescorla-Wagner theory (Equation 8). Similarly, during the second link of the schedule, μ_3 is updated because μ_2 is constant while V_2, the reinforcing value of $S^D{}_2$, is being updated.

In our theory, Equation 9 synthesizes the mechanics of reinforcement learning and the mechanics of conditioned reinforcement. It states that the selection pressure exerted by a discriminative stimulus in a link of a chained schedule is a function of the reinforcing value of that stimulus as determined by the Rescorla-Wagner equation.

4. CONCLUSION

Our biologically-inspired computational theory of adaptive behavior consists of four components and a synthesizing equation. The four components are a statics and dynamics of reinforcement learning (Equation 1 and the evolutionary algorithm), and a statics and dynamics of conditioned reinforcement (Equations 7 and 8). The synthesizing equation (Equation 9) connects the dynamic components of the theory. The static components describe known properties of the steady-state behavior of live organisms. The dynamic components are theories about how those steady-state properties develop. Computational experiments [8] have shown that the evolutionary reinforcement mechanism, which is the cornerstone of our theory, produces equilibria that are accurately described by Equation 1. Ongoing research in our laboratories indicates that the Rescorla-Wagner dynamics produces equilibria that are at least qualitatively consistent with Equation 7. Their quantitative agreement with the equation is currently being assessed.

The next step in our research program is to study in detail the behavior of an artificial organism that is animated by our theory, as it works on chained schedules of reinforcement.

5. REFERENCES

[1] Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *J. Math. Psych*. 47: 109-121.

[2] Davison, M., and McCarthy, D. (1988). *The matching law*. Hillsdale, NJ: Erlbaum.

[3] Herrnstein, R. (1970). On the law of effect. *J. Exp. Anal. Beh.* 13: 243-266.

[4] Kaelbling, L., Littman, M., and Moore, A. (1996). Reinforcement learning: A survey. *J. Art. Int. Res.* 4: 237-285.

[5] Mazur, J. (1984). Tests of an equivalence rule for fixed and variable reinforcer delays. *J. Exp. Psych.: Anim. Beh. Proc.* 10: 426-436.

[6] Mazur, J. (1997). Choice, delay, probability, and condition-ed reinforcement. *Anim. Learn. Beh*. 25: 131-147.

[7] McDowell, J. (1988). Matching theory in natural human environments. *The Beh. Anal*. 11: 95-109.

[8] McDowell, J. (2004). A computational model of selection by consequences. *J. Exp. Anal. Beh.* 81: 297-317.

[9] McDowell, J. (2005). On the classic and modern theories of matching. *J. Exp. Anal. Beh*. 84: 111-127.

[10] Moriarty, D., Schultz, A., and Grefenstette, J. (1999). Evolutionary algorithms for reinforcement learning. *J. Art. Int. Res.* 11: 241-276.

[11] Pierce, W. and Cheney, C. (2004). *Behavior analysis and learning*. Mahwah, NJ: Erlbaum.

[12] Rachlin, H., Battalio, R., Kagel, J., and Green, L. (1981). Maximization theory in behavioral psychology. *Behav. Brain Sci.* 4: 371-417.

[13] Russell, S. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Saddle River, NJ: Prentice-Hall.

[14] Sutton, R., and Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

[15] Touretzky, D., and Saksida, L. (1997). Operant conditioning in Skinnerbots. *Adap. Beh*. 5: 219-24.